Technical Report TR71, October 2007
Technical Report: OSU-CISRC-10/07-TR71
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210-1277, USA
Web site: **http://www.cse.ohio-state.edu/research/techReport.html**
Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2007**
File in pdf format: **TR71.pdf**

On the Optimality of Ideal Binary Time-Frequency Masks

Yipeng Li* and DeLiang Wang

Abstract

The concept of ideal binary time-frequency masks has received attention recently in monaural and binaural sound separation. Although often assumed, the optimality of ideal binary masks in terms of signal-to-noise ratio has not been rigorously addressed. In this paper we give a formal treatment on this issue and clarify the conditions for ideal binary masks to be optimal. We also experimentally compare the performance of ideal binary masks to that of ideal ratio masks on a speech mixture database and a music database. The results show that ideal binary masks are close in performance to ideal ratio masks which are closely related to the Wiener filter, the theoretically optimal linear filter.

Index Terms

Ideal binary mask, ideal ratio mask, optimality, sound separation, Wiener filter

I. Introduction

Recently monaural and binaural sound separation have received considerable attention. A promising approach to the problem, called *computational auditory scene analysis* (CASA) [25], is inspired by the perceptual theory of *auditory scene analysis* (ASA) [2], which attempts to explain the remarkable capability of the human auditory system in segregating an acoustic signal into streams that correspond to different sound sources. The majority of CASA systems developed so far [3], [24], [20], [15] have applied binary time-frequency (T-F) masking to extracting a target sound. Typically, in such systems a signal is first transformed to a T-F representation such as a spectrogram or a cochleagram. Then an element of such a representation, called a T-F unit

Y. Li* is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210-1277, USA. Phone: 1-614-292-7402; Fax: 1-614-292-2911; Email: liyip@cse.ohio-state.edu.

D. L. Wang is with the Department of Computer Science and Engineering & Center of Cognitive Science, The Ohio State University, Columbus, OH, 43210-1277, USA. Phone: 1-614-292-6827; Fax: 1-614-292-2911; Email: dwang@cse.ohio-state.edu.

corresponding to a certain time and frequency, is assigned 1 if its energy is considered as from the target or 0 otherwise. Hu and Wang [11], [10] proposed a binary mask where a T-F unit is assigned 1 if in that unit the target energy exceeds the interference energy and 0 otherwise. They called such a mask the *ideal binary mask* (IBM) because it represents the computational objective of their system and its construction requires the premixing target and interference signals. The IBM has several desirable properties as the computational goal of CASA systems [23], such as the flexibility of selecting targets and the well-definedness. Human speech intelligibility experiments have also shown that target speech reconstructed from the IBM gives high intelligibility scores, even in very low SNR conditions [20], [1], [4]. Several CASA algorithms that directly estimate the IBM [20], [10] have produced good results in speech separation.

A widely used metric for performance measure in sound separation is signal-to-noise ratio (SNR). For sound separation it is defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_n x^2[n]}{\sum_n (\hat{x}[n] - x[n])^2}, \tag{1}$$

where $x[n]$ is a target signal and $\hat{x}[n]$ is the estimated target signal. It has been noted that the IBM is locally optimal in the SNR sense, i.e., flipping a T-F unit's assignment in the IBM always yields a lower SNR for that T-F unit. It has also been assumed that the ideal binary mask is globally optimal, i.e., the IBM produces an output with the highest SNR gain among *all binary masks* [23]. There exist two arguments for the global optimality of the IBM. Hu and Wang [10] argue for the global optimality based on its local optimality. At each T-F unit, the IBM either maximally retains target energy or removes interference energy. As a result, the sum of missing target energy that is discarded by the mask and interference energy that gets through the mask, i.e., the denominator in (1), is minimized. Therefore the IBM would achieve the highest SNR. This argument is flawed in that SNR calculation is nonlinear: the denominator in (1) is not equal to the linear combination of energy retained or removed in each individual T-F unit. Ellis [7] makes an argument from the viewpoint of Wiener filtering. According to Wiener filtering, optimal SNR can be achieved by the Wiener filter whose frequency response is $P_T/(P_T + P_I)$, where P_T and P_I are the power spectrum densities of target and interference signals, respectively. Quantizing the Wiener filter at each T-F unit to the closest binary value results in the IBM which would produce the optimal binary mask. However, this argument suffers the same drawback as the one by Hu and Wang since it is still based on the local optimality of the IBM: the optimal quantization is performed on each T-F unit.

The concept of the IBM with its assumed global optimality has received increasing attention. A recent psychoacoustic study used the IBM to investigate the factors that affect glimpsing of speech in noise [14]. Many recent computational systems have used the IBM as a measure of ceiling performance for source separation [15], [18], [6], [19], [13], [9]. It is therefore worth examining the optimality of the IBM in a more rigorous way. In this paper we give a formal treatment on the optimality of the IBM. We consider the optimality of the IBM at three levels: the T-F unit level, the time frame level, and the global level, and find that local optimality does not translate to global optimality. In Section II we show that, at each level, the IBM can be optimal under certain conditions imposed on T-F decomposition. We also give counterexamples showing that the IBM is not optimal when these conditions are violated. In Section III we compare SNR gain of the IBM to that of ideal ratio masks which are closely related to the Wiener filter. Conclusion and discussion are presented in Section IV.

II. The optimality of the ideal binary mask at different levels

A. T-F Unit Level

Given a T-F decomposition, we consider $\mathbf{X}_{cm}$ and $\mathbf{Y}_{cm}$, the spectral values of a target signal and an interference signal at T-F unit u_{cm}, respectively. c is the frequency index and m the time frame index. At the T-F unit level, the definition of SNR in (1) should be changed slightly when spectral values instead of time-domain signals are used:

$$\text{SNR} = 10\log_{10}\frac{|\mathbf{X}_{cm}|^2}{|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2}, \tag{2}$$

where $\hat{\mathbf{X}}_{cm}$ is the estimated spectral value of the target. According to the definition of the IBM,

$$\hat{\mathbf{X}}_{cm} = \begin{cases} \mathbf{X}_{cm}+\mathbf{Y}_{cm}, & \text{if } |\mathbf{X}_{cm}|^2 > |\mathbf{Y}_{cm}|^2, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here we assume that the frequency decomposition is linear, which is the case for most sound separation systems. Therefore, when a T-F unit is assigned 1, the spectral estimate of the target is the linear combination of $\mathbf{X}_{cm}$ and $\mathbf{Y}_{cm}$.

Consider the case where $|\mathbf{X}_{cm}|^2 > |\mathbf{Y}_{cm}|^2$, i.e., the target is stronger in energy than the interference at u_{cm}. If u_{cm} is assigned 1 as in the IBM, then the denominator in (2) is

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}+\mathbf{Y}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{Y}_{cm}|^2. \tag{4}$$

On the other hand, if u_{cm} is assigned 0, the denominator is

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |0-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}|^2. \tag{5}$$

Since $|\mathbf{Y}_{cm}|^2 < |\mathbf{X}_{cm}|^2$, the denominator is smaller when u_{cm} is assigned according to the IBM.

Similarly, if $|\mathbf{X}_{cm}|^2 \leqslant |\mathbf{Y}_{cm}|^2$, i.e., the target is no stronger in energy than the interference, according to the IBM, u_{cm} is assigned 0 and the denominator becomes

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |0-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}|^2. \tag{6}$$

If u_{cm} is assigned 1, then

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}+\mathbf{Y}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{Y}_{cm}|^2. \tag{7}$$

Since $|\mathbf{X}_{cm}|^2 \leqslant |\mathbf{Y}_{cm}|^2$, the IBM yields a denominator no greater than its alternative. Based on the above discussion, the IBM always minimizes the denominator and consequently maximizes the SNR. Therefore we conclude that the IBM is optimal at the T-F unit level.

B. Time Frame Level

Now consider $x_m[n]$, the time-domain target signal at frame m. Without loss of generality, we assume that the index of n is from 0 to $N-1$. The discrete Fourier transform (DFT) of $x_m[n]$ is

$$\mathbf{X}_{cm} = \sum_{n=0}^{N-1} x_m[n]e^{-\frac{2\pi cn}{N}j}, \quad c = 0,\ldots,N-1,$$

The SNR of $\hat{x}_m[n]$, the estimate of $x_m[n]$, with respect to $x_m[n]$ can be calculated using (1) with summation of n from 0 to $N-1$. It is clear from (1) that maximizing the SNR is the same as minimizing the denominator, the energy of the error signal $\hat{x}_m[n] - x_m[n]$. According to the Parseval's theorem [17], the energy of the error signal can be equivalently calculated in the frequency domain, i.e.,

$$\sum_{n=0}^{N-1}(\hat{x}_m[n] - x_m[n])^2 = \frac{1}{N}\sum_{c=0}^{N-1}|\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2, \tag{8}$$

In Section II-A, we have shown that the IBM minimizes $|\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2$ for each c. Therefore the IBM also minimizes the summation $\sum_{c=0}^{N-1}|\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2$. As a result, the IBM yields the highest SNR among all binary masks.

The key step in the above proof is applying the Parseval's theorem to equating the energy summation in the time domain to that in the spectral domain. This is possible because the bases used in DFT are orthogonal. This relationship of the energy in the time domain and in some transformed domain can be generalized to any orthogonal decomposition. It can be seen as the following: Let $\{\mathbf{e}_i\}$ a complete set of bases with $\langle \mathbf{e}_i, \mathbf{e}_j\rangle = \delta_{ij}$, where $\langle\cdot,\cdot\rangle$ is the inner product in the Euclidian space and δ_{ij} the Dirac delta function. If $\langle \mathbf{x}, \mathbf{e}_i\rangle = a_i$, the projection of vector $\mathbf{x}$ on basis $\mathbf{e}_i$, then we have

$$\langle \mathbf{x}, \mathbf{x}\rangle = \langle \sum_i a_i\mathbf{e}_i, \sum_j a_j\mathbf{e}_j\rangle = \sum_i\sum_j a_i a_j\langle \mathbf{e}_i, \mathbf{e}_j\rangle = \sum_i a_i^2. \tag{9}$$

Therefore we can conclude that a sufficient condition for the IBM to be optimal at the time frame level is orthogonal frequency decomposition.

C. Global Level

Now consider the entire target signal $x[n]$, which is processed frame by frame. To explore conditions for the IBM to be optimal at the global level, we can write $x[n]$ as

$$x[n] = \sum_{m=0}^{M-1} x_m[n]/A[n], \tag{10}$$

where m is the frame index and M the number of frames. $A[n]$ is the normalization factor and $A[n] = \sum_{m=0}^{M-1} w[n - m\tau]$, where w is a window function with length N and τ is the frame shift. Note now $x_m[n] = 0$ for $n < m\tau$ and $n \geqslant m\tau + N$. Similarly we can write the entire estimated signal as

$$\hat{x}[n] = \sum_{m=0}^{M-1} \hat{x}_m[n]/A[n], \tag{11}$$

Again $\hat{x}_m[n] = 0$ for $n < m\tau$ and $n \geqslant m\tau + N$.

The energy of the entire error signal is

$$\begin{aligned}
&\sum_n (\hat{x}[n]-x[n])^2 \\
=&\sum_n (\sum_m \hat{x}_m[n]/A[n] - \sum_m x_m[n]/A[n])^2 \\
=&\sum_n \frac{1}{A^2[n]} (\sum_m (\hat{x}_m[n]-x_m[n]))^2 \\
=&\sum_n \frac{1}{A^2[n]} (\sum_m (\hat{x}_m[n]-x_m[n])^2 + \\
&2\sum_{m_1}\sum_{m_2>m_1} (\hat{x}_{m_1}[n]-x_{m_1}[n])(\hat{x}_{m_2}[n]-x_{m_2}[n])).
\end{aligned} \tag{12}$$

If consecutive frames do not overlap, for a particular n, either $\hat{x}_{m_1}[n]-x_{m_1}[n]$ or $\hat{x}_{m_2}[n]-x_{m_2}[n]$ is zero. This is because a frame is zero outside of its corresponding window and $m_1 \neq m_2$. In this case, the cross terms in (12) do not contribute to the overall error energy and (12) becomes

$$\sum_n (\hat{x}[n]-x[n])^2 = \sum_n \frac{1}{A^2[n]} \sum_m (\hat{x}_m[n]-x_m[n])^2. \tag{13}$$

Assume $A[n]$ is constant for all n, we have

$$\sum_n (\hat{x}[n]-x[n])^2 = \frac{1}{A^2} \sum_m \sum_n (\hat{x}_m[n]-x_m[n])^2. \tag{14}$$

Note that in the above equation, the order of summation is also switched.

Since the IBM minimizes $\sum_n (\hat{x}_m[n]-x_m[n])^2$ for each frame m as discussed in Section II-B, it also minimizes the energy of the entire error signal. Consequently, the IBM is optimal. Given non-overlapping consecutive frames, the window function must be rectangular in order for $A[n]$ to be constant. Non-overlapping windowing can be considered as an orthogonal decomposition of a signal in the time domain. Therefore we can conclude that a sufficient condition for the IBM to be optimal at the global level is orthogonal T-F decomposition with a rectangular window.

When consecutive frames overlap, the cross terms also contribute to the overall energy of the error signal. To see how flipping the assignment of a T-F unit affect the energy, we consider a frame and its immediate successor that overlaps with it. Specifically, we consider $\sum_n (\hat{x}_{m_1}[n]-x_{m_1}[n])^2 + 2\sum_n (\hat{x}_{m_1}[n]-x_{m_1}[n])(\hat{x}_{m_2}[n]-x_{m_2}[n])$, where $m_2 = m_1 + 1$. Assume at frame m_1, a T-F unit with frequency index c^* is flipped from the IBM assignment of 0 to 1. We denote the resulted time-domain estimate of frame m_1 as $\tilde{x}_{m_1}[n]$. Using DFT for frequency decomposition, we first consider the energy change of the square term:

$$\begin{aligned}
\Delta E_s &= \sum_n (\tilde{x}_{m_1}[n]-x_{m_1}[n])^2 - \sum_n (\hat{x}_{m_1}[n]-x_{m_1}[n])^2 \\
&= \sum_c |\tilde{\mathbf{X}}_{cm_1} - \mathbf{X}_{cm_1}|^2 - \sum_c |\hat{\mathbf{X}}_{cm_1} - \mathbf{X}_{cm_1}|^2 \\
&= |\tilde{\mathbf{X}}_{c^*m_1} - \mathbf{X}_{c^*m_1}|^2 - |\hat{\mathbf{X}}_{c^*m_1} - \mathbf{X}_{c^*m_1}|^2 \\
&= |\mathbf{X}_{c^*m_1} + \mathbf{Y}_{c^*m_1} - \mathbf{X}_{c^*m_1}|^2 - |0 - \mathbf{X}_{c^*m_1}|^2 \\
&= |\mathbf{Y}_{c^*m_1}|^2 - |\mathbf{X}_{c^*m_1}|^2.
\end{aligned} \tag{15}$$

where $\tilde{\mathbf{X}}_{cm_1}$ is the estimated spectrum of $\mathbf{X}_{cm_1}$ with the flipped mask. $\mathbf{Y}_{c^*m_1}$ is the spectral value of the interference signal $y_{m_1}[n]$ at c^*. Since u_{c^*m} is assigned 0 in the IBM, $|\mathbf{Y}_{c^*m_1}|^2 > |\mathbf{X}_{c^*m_1}|^2$. Therefore $\Delta E_s > 0$, indicating an energy increase by the flipping.

Now consider the energy change resulted from the cross term:

$$\begin{aligned}\Delta E_c &= 2\sum_n (\tilde{x}_{m_1}[n] - x_{m_1}[n])(\hat{x}_{m_2}[n] - x_{m_2}[n]) \\ &\quad - 2\sum_n (\hat{x}_{m_1}[n] - x_{m_1}[n])(\hat{x}_{m_2}[n] - x_{m_2}[n]) \\ &= 2\sum_n (\tilde{x}_{m_1}[n] - \hat{x}_{m_1}[n])(\hat{x}_{m_2}[n] - x_{m_2}[n]). \end{aligned} \tag{16}$$

To see the effect of flipping on the cross term, we express the time domain signals in the frequency domain. Since $\hat{x}_{m_1}[n]$ is the inverse DFT of $\hat{\mathbf{X}}_{cm_1}$ shifted to the position of the m_1^{th} frame, it can be written as

$$\hat{x}_{m_1}[n] = \begin{cases} \frac{1}{N}\sum_{c=0}^{N-1} \hat{\mathbf{X}}_{cm_1} e^{\frac{2\pi c(n-m_1\tau)}{N}j}, & n = m_1\tau, m_1\tau+1, \ldots, m_1\tau+N-1, \\ 0, & \text{else.} \end{cases}$$

Similarly,

$$\tilde{x}_{m_1}[n] = \begin{cases} \frac{1}{N}\sum_{c=0}^{N-1} \tilde{\mathbf{X}}_{cm_1} e^{\frac{2\pi c(n-c_1\tau)}{N}j}, & n = m_1\tau, m_1\tau+1, \ldots, m_1\tau+N-1, \\ 0, & \text{else,} \end{cases}$$

$$\hat{x}_{m_2}[n] = \begin{cases} \frac{1}{N}\sum_{c=0}^{N-1} \hat{\mathbf{X}}_{cm_2} e^{\frac{2\pi c(n-m_2\tau)}{N}j}, & n = m_2\tau, m_2\tau+1, \ldots, m_2\tau+N-1, \\ 0, & \text{else.} \end{cases}$$

For $n = m_2\tau, \ldots, m_1\tau + N - 1$, the overlapping part of frame m_1 and m_2,

$$\begin{aligned} &\tilde{x}_{m_1}[n] - \hat{x}_{m_1}[n] \\ =&\frac{1}{N}\sum_c \tilde{\mathbf{X}}_{cm_1} e^{\frac{2\pi c(n-m_1\tau)}{N}j} - \frac{1}{N}\sum_c \hat{\mathbf{X}}_{cm_1} e^{\frac{2\pi c(n-m_1\tau)}{N}j} \\ =&\frac{1}{N}(\mathbf{X}_{c^*m_1} + \mathbf{Y}_{c^*m_1}) e^{\frac{2\pi c^*(n-m_1\tau)}{N}j}. \end{aligned} \tag{17}$$

$\hat{x}_{m_2}[n] - x_{m_2}[n]$ can also be represented in the frequency domain:

$$\begin{aligned} &\hat{x}_{m_2}[n] - x_{m_2}[n] \\ =&\frac{1}{N}\sum_c (\hat{\mathbf{X}}_{cm_2} - \mathbf{X}_{cm_2}) e^{\frac{2\pi c(n-m_2\tau)}{N}j} \end{aligned} \tag{18}$$

Using Equations (17) and (18), the energy change of the cross term can be written as

$$\begin{aligned}\Delta E_c =& \sum_n (\tilde{x}_{m_1}[n] - \hat{x}_{m_1}[n])(\hat{x}_{m_2}[n] - x_{m_2}[n])\\ =& \frac{1}{N^2} \sum_n ((\mathbf{X}_{c^* m_1} + \mathbf{Y}_{c^* m_1}) e^{\frac{2\pi c^*(n-m_1\tau)}{N}j} \sum_c (\hat{\mathbf{X}}_{cm_2} - \mathbf{X}_{cm_2}) e^{\frac{2\pi c(n-m_2\tau)}{N}j})\\ =& \frac{1}{N^2} (\mathbf{X}_{c^* m_1} + \mathbf{Y}_{c^* m_1}) \sum_c (\hat{\mathbf{X}}_{cm_2} - \mathbf{X}_{cm_2}) \sum_n e^{\frac{2\pi c^*(n-m_1\tau)}{N}j} e^{\frac{2\pi c(n-m_2\tau)}{N}j}\\ =& \frac{1}{N^2} (\mathbf{X}_{c^* m_1} + \mathbf{Y}_{c^* m_1}) \sum_c (\hat{\mathbf{X}}_{cm_2} - \mathbf{X}_{cm_2}) Q[c]. \end{aligned} \qquad (19)$$

$Q[c] = \sum_n e^{\frac{2\pi c^*(n-m_1\tau)}{N}j} e^{\frac{2\pi c(n-m_2\tau)}{N}j}$ is a phase term. As can be seen from the above equation, when a T-F unit is flipped, it will contribute to the overall energy by coupling with T-F units in the overlapping frames. This becomes clear when the overlapping is 50%, i.e., $\tau = N/2$. In this case, the phase term is

$$\begin{aligned} Q[c] &= \sum_{n=m_2\tau}^{m_1\tau+N-1} e^{\frac{2\pi c^*(n-m_1\tau)}{N}j} e^{\frac{2\pi c(n-m_2\tau)}{N}j}\\ &= \sum_{n'=0}^{N/2-1} e^{\frac{2\pi(c^*+c)n'}{N}j} e^{\frac{2\pi c(m_2\tau - m_1\tau)}{N}j}\\ &= \sum_{n'=0}^{N/2-1} e^{\frac{2\pi(c^*+c)n'}{N}j} e^{\pi c j}\\ &= (-1)^c \frac{1-(-1)^{c+c^*}}{1-e^{\frac{2\pi(c+c^*)}{N}j}}\\ &= \frac{(-1)^c - (-1)^{c^*}}{1-e^{\frac{2\pi(c+c^*)}{N}j}}\\ &= \begin{cases} 0 & \text{if } c - c^* \bmod 2 = 0,\\ -\frac{2(-1)^{c^*}}{1-e^{\frac{2\pi(c+c^*)}{N}j}} & \text{else,} \end{cases} \end{aligned} \qquad (20)$$

where mod denotes the modulo operation. The above derivation indicates that the flipped T-F unit will couple with every other T-F unit in the overlapping frame. Because of the coupling, it is in general difficult to compare the energy change resulting from the square term and that from the cross term. In other words, the cross term may result in energy decrease that outweighs energy increase from the square term and yield lower error energy. Because of the nonlinearity in the SNR calculation, we suspect that the IBM may not be optimal in the overlapping case. We will show in the next subsection that other binary masks can indeed give higher SNR in this case.

The following summarizes the main analytical result in the form of a theorem:

Theorem 1. *A sufficient condition for the ideal binary mask to be globally optimal is orthogonal time-frequency decomposition with rectangular windowing.*

D. *Counterexamples*

In this section we show several counterexamples in which the IBM is not optimal. Note that it is not difficult to come up with such counterexamples. This suggests that the IBM is probably not optimal when the conditions stated before, i.e., orthogonal T-F decomposition and rectangular windowing, are not satisfied. In all examples, signals are sampled at 20 kHz and are mixed to 0 dB SNR to create mixtures for analysis.

We first present a counterexample showing that the IBM is not optimal when a non-orthogonal gammatone filterbank is used for frequency decomposition. The gammatone filterbank has been widely used in CASA systems for frequency decomposition [25]. The impulse response of a gammatone filter is

$$g[n] = \begin{cases} (nT)^{l-1}\text{exp}(-2\pi bnT)\text{cos}(2\pi fnT), & n \geq 0 \\ 0 & \text{else} \end{cases} \tag{21}$$

where T is the sampling interval, $l = 4$ is the order of the filter, b is the equivalent rectangular bandwidth (ERB), and f is the center frequency of the filter. Typically, a gammatone filterbank consists of 32 to 128 filters with f quasi-logrithmically spaced, based on the ERB-rate. The gammatone filterbank does not provide an orthogonal frequency decomposition of a signal because neighboring gammatone filters overlap, especially in the higher frequency range. Fig. 1 shows an example that the IBM is not optimal with the gammatone filterbank for a single frame. In this example, the gammatone filterbank has 128 channels and the center frequencies are linearly spaced from 50 to 8000 Hz on the ERB-rate scale. The top two panels show two musical signals with 2048 data points. The lower left is the IBM and the lower right is a binary mask obtained with a local SNR threshold (LC) [4] of 1 dB, i.e., u_{cm} is labeled 1 if and only if $10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n(\hat{x}_{cm}[n]-x_{cm}[n])^2} > 1$, where $x_{cm}[n]$ is the time-domain signal underlying u_{cm} and $\hat{x}_{cm}[n]$ is the estimate. In all the illustrations in this subsection, white indicates that a T-F unit is labeled 1 and black 0. The estimated signals are reconstructed from the two binary masks using a technique introduced by Weintraub [26] (also see [25]). Since the resynthesis procedure is an integrated part of the gammatone filterbank-based analysis, we do not attempt to isolate its contribution to the SNR gain. In this case, the IBM gives a 7.0 dB SNR gain while the other binary mask gives a 7.3 dB SNR gain.

The second counterexample, illustrated in Fig. 2, shows that the IBM is not optimal when a non-rectangular window is used with orthogonal T-F decomposition. In this example, the DFT is used for frequency decomposition and the frames do not overlap. The top two panels plot two musical signals. When a hamming window with a length of 512 samples is applied, the SNR gain of the IBM (lower left) is 3.97 dB while the SNR gain of a mask (lower right) with a LC of 0.4 dB is 3.99 dB. One of the noticeable differences between the two masks is indicated by a circle.

If consecutive frames overlap, the IBM may not be optimal even with a rectangular window. Fig. 3 shows such an example with the same musical signals as in Fig. 2. The frame length is 512 and the overlapping is 50%. The SNR gain of the IBM (lower left) is 16.7 dB while the SNR gain for a mask obtained with a LC of 0.4 dB is 16.9 dB (lower right). The circle marks one noticeable difference between the two masks. Recalling equations (15) and (16), this example shows the effect of coupling between overlapping frames.

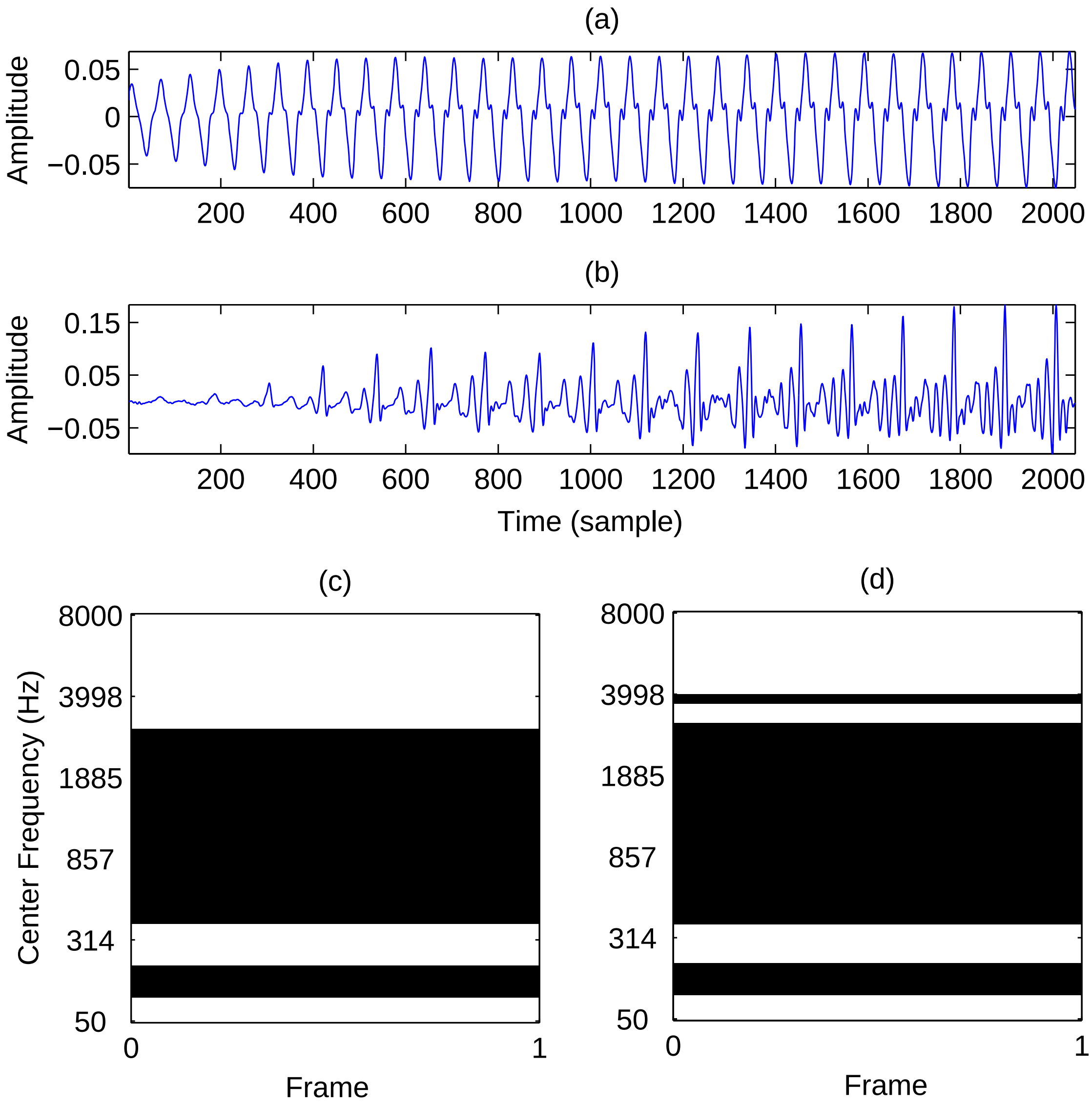

Fig. 1. An example showing that the IBM is not optimal when a gammatone filterbank is used for frequency decomposition for one frame. (a). The waveform of a target music signal. (b). The waveform of an interference music signal. (c). The IBM. (d) A mask generated with a LC of 1 dB.

III. The ideal binary mask and the ideal ratio mask

Most sound separation systems decompose a signal into overlapping frames to reduce boundary effects caused by windowing. In this case, based on the discussion in Sections II-C and II-D, the IBM may be optimal. On the other hand, its SNR gain is close to that of ideal ratio masks (IRM). The IRM is defined as [21]

$$R_{c,m} = \frac{|\mathbf{X}_{cm}|^2}{|\mathbf{X}_{cm}|^2 + |\mathbf{Y}_{cm}|^2} \tag{22}$$

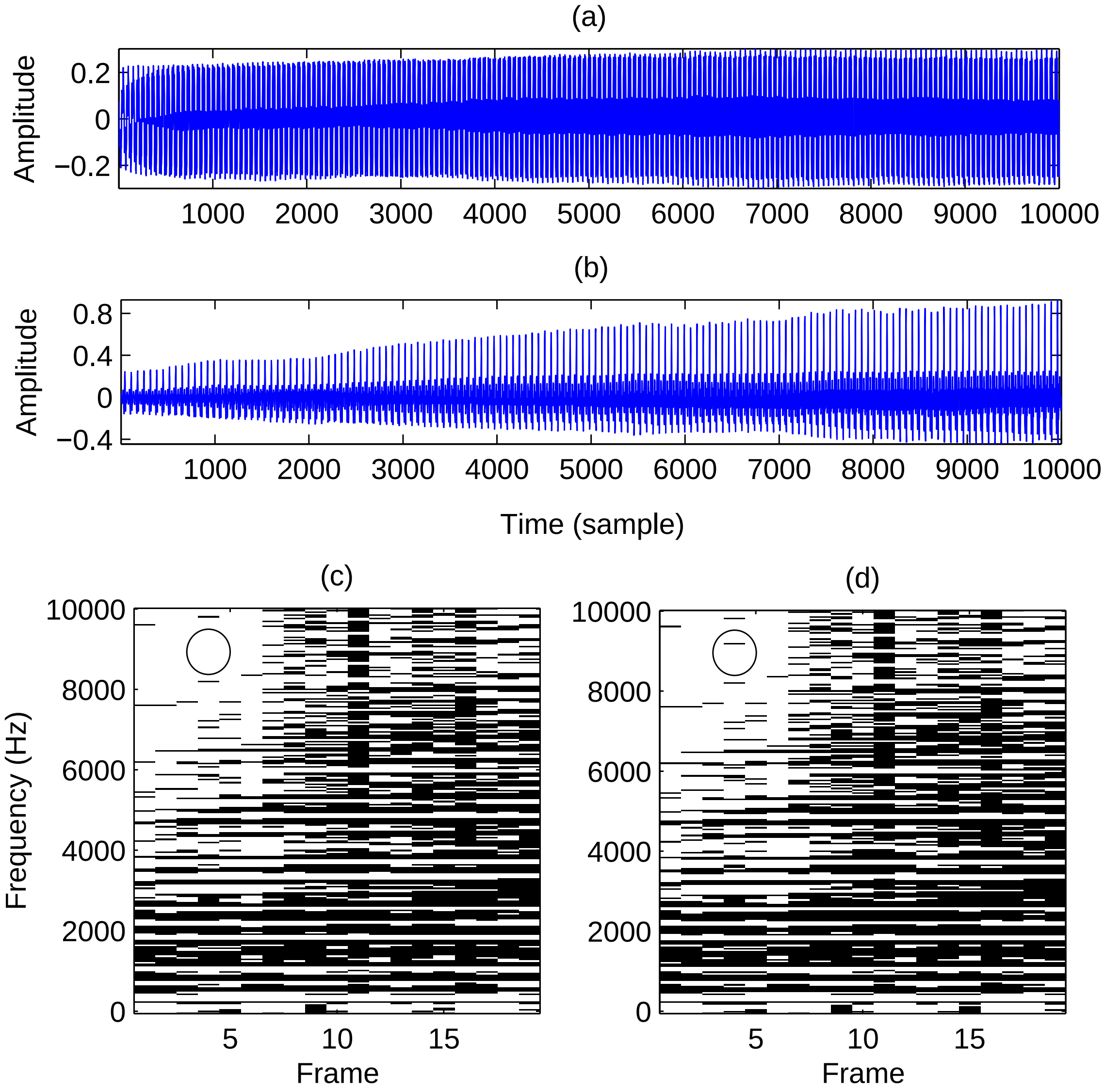

Fig. 2. An example showing that the IBM is not optimal when a hamming window is used for orthogonal T-F decomposition. (a). The waveform of a target music signal. (b). The waveform of an interference music signal. (c). The IBM. (d). A mask generated with a LC of 0.4 dB.

for each c and m. The IRM is closely related to the Wiener filter, the optimal linear filter in the minimum mean-square error sense [27]. Moreover, if a target signal, an interference signal, and their mixture are jointly Gaussian, the Wiener filter is the optimal filter among all possible filters, linear or nonlinear [22]. Additionally, given that the causality of a filter is not required and the target signal and the interference signal are uncorrelated, the Wiener filter amounts to the same ratio as (22) with spectral values replaced by power spectral densities [22]. The conditions for the Wiener filter to be a ratio mask are satisfied in most cases: the non-causality of the filter can be allowed since most sound separation systems operate offline; the uncorrelatedness can also

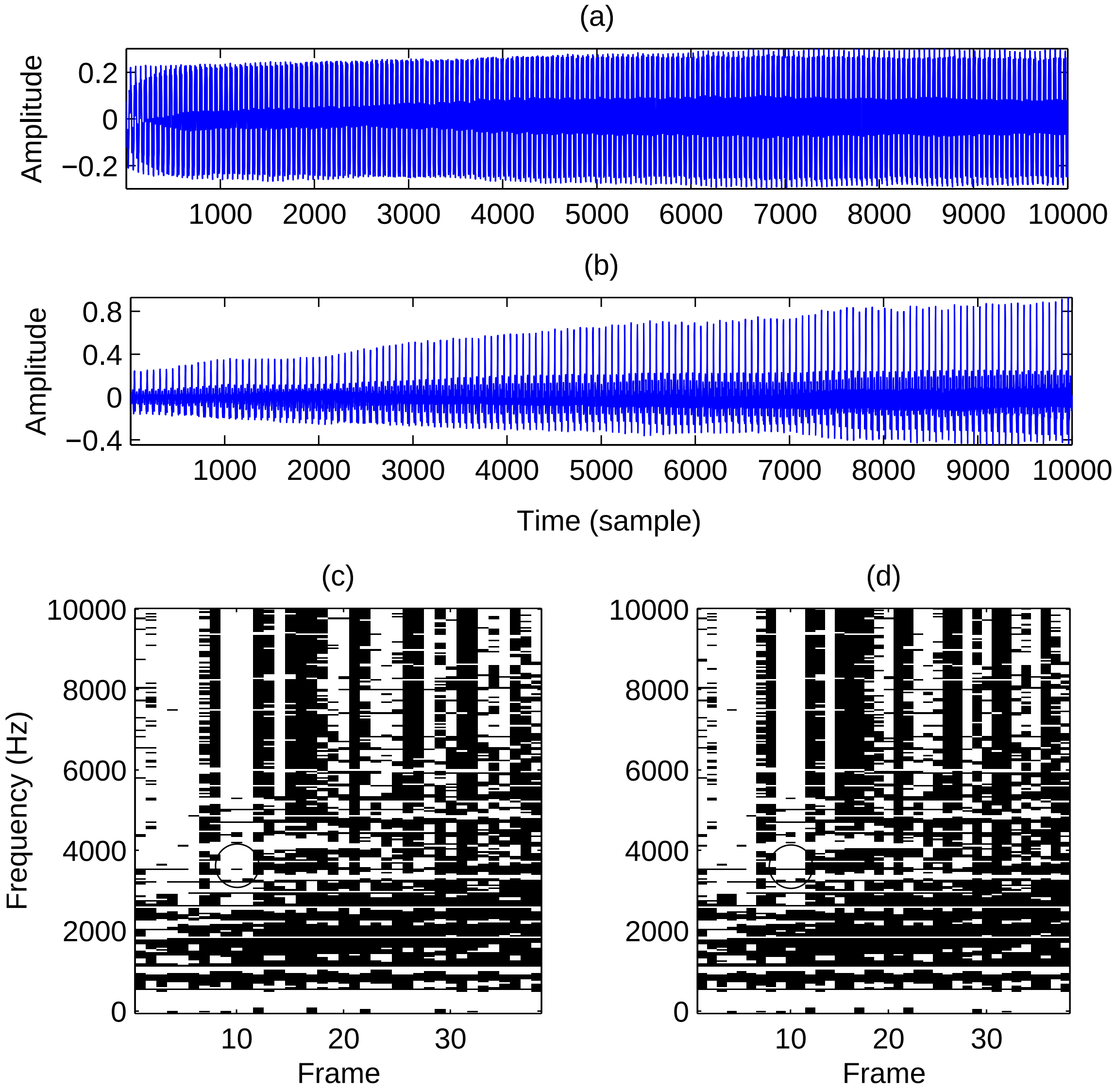

Fig. 3. An example showing that the IBM is not optimal when frames overlap even with a rectangular window. (a). The waveform of a target music signal. (b). The waveform of an interference music signal. (c). The IBM. (d). A mask generated with a LC of 0.4 dB.

be assumed since sound sources are generally independent.

One can show that the IRM always leads to a local SNR gain no smaller than the IBM. For u_{cm}, we can define three underlying signals: the target $x_{cm}[n]$, the interference $y_{cm}[n]$, and the mixture $z_{cm}[n]$. With linear frequency decomposition, $z_{cm}[n] = x_{cm}[n] + y_{cm}[n]$. The ratio mask can be defined using the energy of time-domain signals as

$$r = \frac{\sum_n x^2_{cm}[n]}{\sum_n x^2_{cm}[n] + \sum_n y^2_{cm}[n]}. \tag{23}$$

Denote $E = \sum_n x_{cm}^2[n] + \sum_n y_{cm}^2[n]$ and we have $\sum_n x_{cm}^2[n] = rE$ and $\sum_n y_{cm}^2[n] = (1-r)E$. Consider the case where $\sum_n x_{cm}^2[n] > \sum_n y_{cm}^2[n]$, the target stronger than the interference. According to the IBM, $z_{cm}[n]$ is retained and the local SNR is

$$\text{SNR} = 10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n (z_{cm}[n] - x_{cm}[n]])^2}. \tag{24}$$

When the IRM is applied, the new SNR is

$$\text{SNR}' = 10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n (rz_{cm}[n] - x_{cm}[n])^2}. \tag{25}$$

Now compare the denominators in (24) and (25):

$$\begin{aligned}&\sum_n (z_{cm}[n] - x_{cm}[n])^2 - \sum_n (rz_{cm}[n] - x_{cm}[n])^2\\ =&\sum_n y_{cm}^2[n] - \sum_n (ry_{cm}[n] + (r-1)x_{cm}[n])^2\\ =&(1-r^2)\sum_n y_{cm}^2[n] - 2r(r-1)\sum_n x_{cm}[n]y_{cm}[n] - (r-1)^2\sum_n x_{cm}^2[n]\\ =&(1-r^2)(1-r)E - (r-1)^2 rE\\ =&(r-1)^2E. \end{aligned} \tag{26}$$

Note that in the above derivation we assume that the target and the interference are uncorrelated at u_{cm}, i.e., $\sum_n x_{cm}[n]y_{cm}[n] = 0$. Since $(r-1)^2E \geqslant 0$, $\sum_n (z_{cm}[n] - x_{cm}[n])^2 \geqslant \sum_n (rz_{cm}[n] - x_{cm}[n])^2$. This shows that compared to IBM, IRM gives an equal or smaller denominator and therefore the same or better SNR. The equal sign holds when $r = 1$, i.e., when the interference is absent at u_{cm}. Similarly we can show that when $\sum_n x_{cm}^2[n] \leqslant \sum_n y_{cm}^2[n]$, the IRM also achieves an SNR that is at least as good as the IBM. In this case, the equal sign holds when $r = 0$, i.e., when the target is absent at u_{cm}.

In the above discussion, we show that the IRM is locally no worse in terms of SNR compared to the IBM. However it is difficult to theoretically quantify the global difference between the two. We investigate this issue experimentally using mixtures of interest. In particular, we use a speech mixture database and a music database. The speech mixture database is collected by Cooke [5], which includes different types of interference that are commonly encountered in real environment. It also has premixing target and interference, which makes the construction of the IBM and the IRM possible. For music, we use a database constructed for musical sound separation. The database is synthesized from the tenor and the alto line of string quartets by J. S. Bach. Each line is constructed based on MIDI data using instrument samples from the RWC database [8]. Details of synthesis can be found in [16]. For each database, we consider the SNR gain of the IBM and the IRM over two different kinds of frequency decomposition—DFT and the gammatone filterbank (GF) as described in Section II-D. In each case, the frame length is 512 points and the frame overlapping is 50%. The sampling frequency is 20 kHz. In the gammatone filterbank analysis, the filterbank has 64 filters with center frequencies equally spaced on the ERB-rate scale from 50 to 8000 Hz.

Table I shows the SNR gains of the IBM and the IRM in dB for the speech mixture database with both DFT and GF. There are 10 different types of interferences: N0, 1-kHz pure tone; N1,

Interference	DFT		GF	
	IBM	IRM	IBM	IRM
N0	18.3	19.0	21.4	21.7
N1	12.2	12.9	11.3	12.0
N2	17.6	18.3	17.6	18.1
N3	7.8	8.5	7.6	7.8
N4	12.4	13.0	11.2	11.6
N5	18.7	19.4	19.7	19.9
N6	21.0	21.7	20.6	20.9
N7	13.9	14.7	12.4	12.8
N8	13.2	13.8	12.2	12.7
N9	9.7	10.4	9.9	10.1
Average	14.5	15.2	14.4	14.8

TABLE I
SNR GAIN (IN DECIBEL) OF IBM AND IRM FOR A SPEECH MIXTURE DATABASE

white noise; N2, noise bursts; N3, "cocktail party" noise; N4, rock music; N5, siren; N6, trill telephone; N7, female speech; N8, male speech; and N9, female speech. Each number in Table I represents an average over 10 utterances for one type of interference. The average SNR gain over the whole database is listed in the bottom row of the table. It can be seen that the IRM gives a higher SNR gain in all cases. On the other hand, the SNR gains of the IBM are close to those of the IRM. When DFT is used for frequency decomposition, the SNR gain of the IBM is 0.7 dB lower than that of the IRM on average. With the gammatone filterbank, the difference is only 0.4 dB. The variance of the SNR gain difference is also small—the largest difference is 0.8 dB when the interference is female speech (N7) and DFT is used.

The SNR gains for the music database are shown in Table II. In this case, we group the SNR gains according to instrument combinations. Four instruments, a clarinet (CL), a flute (FL), a violin (VN), and a trumpet (TR) are used to synthesize different music lines in the music database and there are 6 different combinations. It can be seen that the IRM gives higher SNR gains for all instrument combinations when DFT is used for frequency decomposition. For the gammatone filterbank, the IBM actually performs better in several instrument combinations. One possible reason is that the uncorrelatedness assumption does not hold well in music. In Western music, pitches in harmonic relation—pitches form a simple integer ratio [12]—are favored. As a result, harmonics of different notes may collide. This is more likely with the gammatone filterbank since the bandwidth of the filters are wider in the high frequency range. Nonetheless, on average, the IRM gives a better SNR gain than that of the IBM. Similar to speech, the SNR gains between the IBM and the IRM are small. With DFT, the IBM is 0.8 dB worse while with the gammatone filterbank, the difference is only 0.1 dB.

In summary, the IRM achieves higher SNR gains compared to the IBM. However, despite the fact that the IBM is binary and the IRM is not, the SNR gain of the IBM is close to that of the IRM. This shows that, although the IBM is not optimal, it still gives a very reasonable performance metric for sound separation.

IV. Concluding Remarks

In this paper we have addressed the optimality of the IBM in terms of SNR gain at three different levels and clarified the conditions at each level for the IBM to be optimal. At the T-F

Instruments	**DFT**		**GF**	
	IBM	IRM	IBM	IRM
CL+FL	12.9	13.5	12.3	12.0
CL+VN	13.2	13.9	12.3	12.1
CL+TR	11.3	12.2	9.0	9.3
FL+VN	13.7	14.8	11.9	11.8
FL+TR	11.1	12.1	8.8	9.4
VN+TR	12.1	12.8	8.9	9.2
Average	12.4	13.2	10.5	10.6

TABLE II
SNR GAIN (IN DECIBEL) OF IBM AND IRM FOR A MUSIC DATABASE

unit level, the IBM is optimal. At the time frame level, the IBM is optimal when the frequency decomposition is orthogonal. At the global level, IBM is optimal when the T-F decomposition is orthogonal and the windowing function is rectangular. We give counterexamples where the IBM is not optimal when the stated conditions are not satisfied. Since in most practical applications, frames overlap, and as a result the IBM is not expected to be optimal. However we have shown experimentally that the performance of the IBM is close to that of the IRM and therefore the IBM is still a good objective for sound separation systems. Note that IBM estimation, unlike IRM estimation, requires only binary decisions, which makes applicable a host of classification and clustering methods.

In our discussion of IBM optimality, we treat a signal deterministically, i.e., we do not consider the statistical properties of a signal, such as stationarity or distribution. We believe our treatment is more appropriate because it makes no assumption about signals. For example, if signals are treated statistically, the energy of the error signal has to be replaced by the expectation of the energy. In this case, we have $E(\sum_n (x_m[n] - \hat{x}_m[n])^2) = E(\sum_c |\mathbf{X}_{cm} - \hat{\mathbf{X}}_{cm}|^2)$. To proceed, i.e., to switch the expectation and summation, one has to assume that error terms of different c are statistically independent. However such an assumption is difficult to justify.

ACKNOWLEDGMENT

This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an AFRL grant (FA8750-04-1-0093).

REFERENCES

[1] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, “Determination of the potential benefit of time-frequency gain manipulation,” *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[2] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[4] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, “Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask,” *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[5] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U. K.: Cambridge University Press, 1993.

[6] O. M. Deshmukh and C. Y. Espy-Wilson, “Speech enhancement using the modified phase-opponency model,” *Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3886–3898, 2007.

[7] D. P. W. Ellis, “Model-based scene analysis,” in *Computational Auditory Scene Analysis: Principles, Algorithms, and Application*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, 2003.

[9] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[10] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[11] ——, "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[12] T. L. Hubbard and D. L. Datteri, "Recognizing the component tones of a major chord," *American Journal of Psychology*, vol. 114, no. 4, pp. 569–589, 2001.

[13] Y.-I. Kim, S. J. An, and R. M. Kil, "Zero-crossing based time-frequency masking for sound segregation," *Neural Information Processing - Letter & Review*, vol. 10, pp. 125–134, 2006.

[14] N. Li and P. C. Loizou, "Factors influencing glimpsing of speech in noise," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1165–1172, 2007.

[15] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2014–2023, 2006.

[16] Y. Li and D. L. Wang, "Pitch detection in polyphonic music using instrument tone models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. II.481–484.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[18] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. Article ID 84 186, 15 pages, 2007, doi:10.1155/2007/84186.

[19] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1766–1776, 2007.

[20] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[21] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[22] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[23] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, pp. 181–197.

[24] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[25] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press, 2006.

[26] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University, Department of Electrical Engineering, 1985.

[27] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press, 1949.